American Journal of Epidemiology
Vol. 171, No. 3
DOI: 10.1093/aje/kwp403
Advance Access publication:
December 31, 2009

Practice of Epidemiology

Fitting General Relative Risk Models for Survival Time and Matched Case-Control Analysis

Bryan Langholz* and David B. Richardson

* Correspondence to Dr. Bryan Langholz, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, 1540 Alcazar Street, CHP-220, Los Angeles, CA 90033 (e-mail: langholz@usc.edu).

Initially submitted May 5, 2009; accepted for publication November 11, 2009.

Cox proportional hazards regression analysis of survival data and conditional logistic regression analysis of matched case-control data are methods that are widely used by epidemiologists. Standard statistical software packages accommodate only log-linear model forms, which imply exponential exposure-response functions and multiplicative interactions. In this paper, the authors describe methods for fitting non-log-linear Cox and conditional logistic regression models. The authors use data from a study of lung cancer mortality among Colorado Plateau uranium miners (1950–1982) to illustrate these methods for fitting general relative risk models to matched case-control control data, countermatched data with weights, *d*:*m* matching, and full cohort Cox regression using the SAS statistical package (SAS Institute Inc., Cary, North Carolina).

algorithms; cohort studies; conditional likelihood; dose-response function; linear trend; logistic models; models, statistical; software

Abbreviations: CI, confidence interval; ERR, excess relative risk; WLM, working level months.

Contemporary approaches to analysis of cohort and case-control data often follow from risk-set sampling designs. For cohort studies, risk-set sampling designs are related to the Cox proportional hazards model; at each failure time, a risk set is formed that includes the index case and a set of controls comprising all other cohort members who are at risk at that time (1). Similarly, contemporary approaches to case-control designs typically involve a sampled risk set, which is a subset of the full risk set enumerated at each failure time (2, 3). These approaches both result in a data structure that looks like matched case-control data with 1 case per case-control set. This motivates the use of the conditional logistic likelihood for analysis of such data (1).

The procedures available in most standard statistical packages, such as SAS (SAS Institute Inc., Cary, North Carolina), Stata (Stata Corporation, College Station, Texas), and R (R Foundation for Statistical Computing, Vienna, Austria), for fitting Cox or conditional logistic regression models limit a data analyst to log-linear models of the form $\log(\varphi(z;\beta)) = z\beta$, where z is a vector of explanatory variables and φ is the rate (or odds) ratio. This implies exponential dose-response trends and multiplicative interactions. The use of model forms other than the log-linear are desirable when they provide a better representation of the exposure response or when they address biologic or public health questions. They also may facilitate exposure-response analyses, since misspecification of model form can lead to loss of power for a model-based test of an exposure-response association and the possibility that estimates of effect for extreme exposure levels will be substantially distorted (4). In recent papers, investigators have described how standard statistical software packages may be used to fit general relative risk models (i.e., relative risk models not constrained to be log-linear in form) in the context of Poisson regression and unconditional logistic regression analyses (5, 6). However, the prior literature has not addressed the fitting of general relative risk models in the context of the widely used approaches of Cox regression analysis of survival-time data or conditional logistic regression analysis of matched case-control data.

In this paper, we describe a relatively simple approach for obtaining maximum conditional logistic likelihood estimates from analysis of case-control and cohort data for general relative risk model forms and illustrate the implementation of this approach using the SAS statistical package.

MATERIALS AND METHODS

Consider a regression model $\varphi(z;\beta)$, where z represents explanatory data from the epidemiologic study, β are parameters to be estimated, and φ denotes the odds or rate ratio function. Through the use of maximum likelihood methods to fit general relative risk models, the φ function can encompass a wide range of models besides the log-linear. The linear excess relative risk (ERR) model, $\varphi(z;\beta) = 1 + z\beta$, is one model form of interest, particularly in environmental and occupational epidemiology. More broadly, Thomas (7) described a class of relative risk models that permit a data analyst to compare linear and log-linear dose-response models or additive and multiplicative interaction models. Models that are a mixture of linear and log-linear forms may facilitate comparison of linear and log-linear models by incorporating each as a special case of the broader set of models under consideration (7, 8). One form of a mixture model is $\varphi(z;\beta) = \left(e^{z\beta}\right)^{\alpha}(1+z\beta)^{1-\alpha}$ (7, 8).

Generally, for any specified φ, the conditional logistic likelihood contribution can be written as

$$L(\beta) = \varphi_{\text{case}}(\beta) / \sum_{\text{case and controls}} \varphi_j(\beta), \qquad (1)$$

where, with Z_j the covariates for case-control subject j, $\varphi_j(\beta) = \varphi\left(Z_j;\beta\right)$.

In this paper, we focus on the use of SAS PROC NLP to fit these models. The procedure computes exact derivatives of the log-likelihood function and produces likelihood estimates via a ridge Newton-Raphson fitting algorithm; other options for optimization are also available. Other statistical packages that provide general optimization routines may be used in the same way to implement the methods.

The analytical data structure for this implementation differs from that employed in an unconditional logistic regression analysis. A data structure that includes 1 record per person under study might be referred to as a person-level data structure (9). In contrast, we propose to create a data structure that includes 1 record per case-control set (i.e., 1 record for a case and its complement of matched controls). We refer to the latter as a case-control data structure.

Consider a case-control study in which incident cases of disease have been ascertained over a period of follow-up. For each case, 2 controls are selected from the study base, defined as the population at risk at the time of case failure. For simplicity, let us assume that the person-level data set includes a binary indicator of case status, *case*, a single explanatory variable of interest, z, and a numeric indicator of matched risk sets, *setno*. A case-control data structure can be generated for the purposes of conditional logistic regression analysis consisting of 1 record for the case and associated controls, with explanatory information being represented by the variables $z1$, $z2$, and $z3$, where $z1$ denotes the case's exposure to z and $z2$ and $z3$ denote the controls' exposures to z. Table 1 illustrates the person-level data structure and the case-control data structure for a hypothetical 1:2 matched case-control study with 7 cases. The case-control data structure shown in Table 1 can be easily generated from the person-level data structure via SAS (Appendix 1).

Table 1. Comparison of a Person-Level Data Structure (plevel) and a Case-Control-Level Data Structure (cclevel) in a Hypothetical 1:2 Matched Case-Control Study With 7 Cases

Person-Level Data Structure			Case-Control-Level Data Structure			
setno	*case*	*z*	*setno*	*z1*	*z2*	*z3*
1	1	1,819.9	1	1,819.9	105.0	1,244.4
1	0	105.0	2	2,367.0	626.9	271.0
1	0	1,244.4	3	2,978.0	281.0	440.2
2	1	2,367.0	4	1,200.8	99.0	355.3
2	0	626.9	5	1,207.0	443.2	196.8
2	0	271.0	6	299.0	158.0	287.0
3	1	2,978.0	7	1,091.1	518.9	202.3
3	0	281.0				
3	0	440.2				
4	1	1,200.8				
4	0	99.0				
4	0	355.3				
5	1	1,207.0				
5	0	443.2				
5	0	196.8				
6	1	299.0				
6	0	158.0				
6	0	287.0				
7	1	1,091.1				
7	0	518.9				
7	0	202.3				

A conditional logistic regression analysis of the association between z and the outcome defining the case series, employing the standard log-linear model form $\varphi\left(z;\beta\right) = e^{(z\beta)}$, may be fitted via SAS as follows:

```
proc nlp data= ;
  parms beta=0;
  L =log(exp(z1*beta) / (exp(z1*beta)+ exp(z2*beta)+
  exp(z3*beta)));
  max L; run;
```

The parms statement tells SAS that β is to be estimated and sets the initial value to 0. The next line computes L, the conditional logistic log-likelihood (equation 1). Finally, the max statement specifies that the log-likelihood is to be maximized with respect to the parameters in the parms statement.

Now let us consider a non-log-linear model, such as the linear ERR model given by $\varphi(z;\beta) = (1 + \beta z)$. This model may be fitted via SAS as follows:

```
proc nlp data= ;
  parms beta=0;
  profile beta / alpha=.05;
  L=log((1+z1*beta)/((1+z1*beta)+(1+z2*beta)+(1+
     z3*beta)));
  max L; run;
```

A likelihood ratio test of $\beta = 0$ can be obtained as twice the (absolute) difference between the log-likelihood at $\beta = 0$ and the log-likelihood at the maximum likelihood estimate in the usual way. In addition, we have included the profile statement, which requests a 95% profile likelihood confidence interval for the β estimate. This distribution of estimated parameters in non-log-linear models is often not symmetrical, and profile likelihood confidence intervals are more accurate than Wald-type intervals.

1:*m* and 1:variable matching

The examples above concern a 1:2 matched case-control study. The approach developed above easily extends from 1:2 matching to 1:*m* matching or 1:variable matching. The latter is of practical importance, since often in case-control studies a full complement of controls is not obtained for all cases. As in the examples above for a 1:2 matched case-control study, a data set is produced that consists of 1 record per risk set, with a vector, *Z*, that indexes explanatory information for the case and its affiliated controls. To accommodate variable matching, a variable *ntot* is included that gives the number of subjects (cases and controls) in each set. For example, consider a case-control study in which up to 10 controls per case were selected. If the case-control set includes the full complement of controls, then *ntot* would equal 11. A conditional logistic regression analysis of these data employing the log-linear model form $\varphi(z;\beta) = e^{(z\beta)}$could be fitted via SAS as follows:

```
proc nlp data=;
  parms beta=0;
  array z[11] z1-z11;
  sum=0; ntot=ntot;
  do i = ntot to 1 by -1;
    phi = exp(z[i]*beta);
    sum = sum + phi;
  end;
  L=log(phi/sum);
  max L; run;
```

Similarly, non-log-linear models, such as a mixture model of the form $\varphi(z;\beta) = \left(e^{z\beta}\right)^{\alpha}(1+z\beta)^{1-\alpha}$, can be easily fitted to 1:variable matched case-control data via SAS (Appendix 2).

Analysis of survival time in the full cohort

Note that the conditional logistic likelihood contribution for a case-control set in equation 1 is identical to the expression for the Cox likelihood contribution from a risk set when there is a single failure at each failure time. As in the Cox likelihood, the numerator of the term is the (relative) hazard for the subject who experienced the event at the index time, and the denominator is the sum of the (relative) hazards for all subjects at risk at the index time (10). Therefore, by including all controls for each case in a risk set, this approach can accommodate Cox proportional hazards regression. Further, stratified Cox regression can be fitted from "matched risk sets" by restricting the risk set to include only the controls who match the case on the stratification variables. Richardson (11) provides a simple SAS macro for enumerating matched risk sets. An analysis employing the log-linear model form $\varphi(z;\beta) = e^{(z\beta)}$yields the standard log-linear Cox regression analysis. However, non-log-linear Cox regression models can be fitted using the approach described above, with the Cox regression analysis implemented as a form of a 1:variable matched case-control analysis.

Sampling weights

In some case-control studies, controls are matched to the index case on a variety of factors and countermatched on exposure status (12). In such a study, investigators need to incorporate sampling weights into the analysis to accommodate the countermatched design (13). In this situation, a weight is constructed for each subject in the case-control set. The weights are set to the countermatching sampling weights. Appendix 3 illustrates extension of this approach to accommodate countermatching sampling weights. In general, the conditional logistic regression of case-control data with "complex" sampling of cases and/or controls can require sampling weights for valid estimation. When sampling is independent over case-control sets, the standard errors and confidence limits from the weighted likelihood are valid (14–16). However, for other sampling designs, such as case-cohort-type designs, variance adjustment may be needed (17–20).

d:*m* matching

The occurrence of 2 or more events at the same point in time is referred to as tied data. In Cox regression analysis, there are several different methods for handling tied data. One approach is Breslow's method (21), which is a standard formula for analysis of *d*:*m* matched data with tied failure times in Cox regression and is valid when the proportion of cases in each risk (or case-control) set is very small. An example of SAS NLP code to fit the Breslow likelihood is available at a University of Southern California Web site (http://hydra.usc.edu/timefactors/examples/example.html). Breslow's method can perform poorly when the proportion of cases in the risk (case-control) sets is not "very small" (22), and we would argue that this approach need not be used. This problem is remedied by use of exact methods; one exact approach to analysis of risk sets with tied data assumes that time is measured in discrete intervals and tied events happened during the same interval. The form of the (standard) conditional logistic likelihood for matched case-control data, accommodating multiple cases in the case-control sets, can be written as

$$L(\beta) = \varphi_D(\beta) \Big/ \sum_{s \subset R : |s| = |D|} \varphi_s(\beta), \qquad (2)$$

where $|s|$ is the number of elements in the set s so that the sum in the denominator is over all subsets of the case-control set of size the number of cases. The number of terms in the denominator increases exponentially with both the number of cases and the number of controls; however, a recursive fitting algorithm can be used to reduce the number of computations to a linear function of cases and controls (1, 23). Appendix 4 provides SAS NLP code that illustrates the use of the recursive algorithm to fit the conditional logistic likelihood given in equation 2, providing an exact method for analysis of tied data.

Empirical example

To illustrate this approach, we use data from a study of underground uranium miners (24). The Colorado Plateau cohort included male workers employed in underground uranium mining operations between January 1, 1950, and December 31, 1960. Vital status was ascertained through December 31, 1982. The outcome of interest, lung cancer mortality, was defined on the basis of underlying cause of death, coded according to the revision of the *International Classification of Diseases* that was in effect at the time of death. The primary exposure of interest was defined as cumulative radon exposure, expressed in working level months (WLM), and was computed for each worker as the product of the length of employment in each job in a year and the estimated radon exposure rate for that job. For each lung cancer death, a risk set was formed that included all workers who were alive and eligible for study inclusion at the age of death of the index case; controls were also matched to cases on calendar year at risk (defined in 5-year categories from <1960 to ≥1990). For example, if the case died of lung cancer at age 62 years in 1958, the controls in the risk set were all cohort members who were aged 62 years between 1955 and 1960 and were in the study at that age. For analysis, we computed cumulative exposure up to 2 years prior to the age of the risk-set index case (15).

First, we illustrate a 1:2 nested case-control study; this data set included 263 lung cancer deaths. Two controls were selected for each lung cancer death by random sampling without replacement from all controls from the risk set. Second, we illustrate a 1:variable matched case-control study, using the nested case-control data, similar to the nested case-control data set described by Langholz et al. (15). Forty controls were selected for each lung cancer death by random sampling without replacement from all controls from the risk set, unless there were fewer than 40 controls in the risk set, in which case all controls were taken. Following the method of Langholz et al. (15), we fitted a linear excess rate ratio model. Third, we illustrate a 1:3 matched study with countermatching. In addition to modeling the effect of cumulative radon exposure, we adjust for cumulative pack-years of cigarette smoking. Fourth, we illustrate proportional hazards regression analysis for full cohort data, again fitting the model with cumulative radon and smoking. In this example, we used cohort data with tied failure times randomly broken so that there was only 1 case in each risk set. Using the conditional logistic likelihood method with multiple cases per set and an exact method to handle tied data, we fitted this same model to the full cohort data with multiple cases in each risk set at a tied failure time. All analyses were conducted using the SAS statistical package (version 9.2) (25). Profile likelihood confidence intervals are reported via the profile statement. The example data sets and SAS code used to perform these analyses are available at http://hydra.usc.edu/timefactors/examples/example.html (topic 10).

RESULTS

We commenced with a 1:2 matched case-control study of the Colorado Plateau miners' cohort. A log-linear conditional logistic regression model was fitted via SAS PROC NLP. The point estimate and standard error for the association between cumulative exposure and lung cancer ($\beta = 0.0453/100$ WLM; standard error, 0.0066) were identical (at 4 decimal places) to the conditional logistic regression result obtained via SAS PROC LOGISTIC.

Next, we fitted a mixture model of the form $\varphi(z;\beta) = \left(e^{z\beta}\right)^{\alpha}(1+z\beta)^{1-\alpha}$ via PROC NLP. A point estimate for the "mixture" parameter, α, close to 1 indicates that the data are more consistent with the log-linear model, while an estimated α close to 0 is more consistent with the ERR model. The point estimate for α ($\alpha = -0.02$, 95% confidence interval (CI): −0.05, 0.00) provided evidence in favor of the ERR model. The mixture model ($-2 \log L = 486.1$) fitted much better than did the log-linear model ($-2 \log L = 501.3$; likelihood ratio $= 501.3 - 468.1 = 33.2$; $P < 0.001$). On the other hand, the mixture model did not fit better than the ERR model ($-2 \log L = 471.4$; likelihood ratio $= 471.4 - 468.1 = 3.3$; $P = 0.07$), so the ERR form was a good choice. Fitting the ERR model for lifetime cumulative exposure yielded an estimated ERR/100 WLM of 0.38 (95% CI: 0.18, 0.95). A test of the null hypothesis yielded a 1-df chi-squared value of 95.4 ($P < 0.001$).

We next fitted a 1:variable matched study in which up to 40 controls per case were sampled from the cohort 5-year calendar-period matched risk sets. Fitting an ERR model for cumulative exposure yielded an estimated ERR/100 WLM of 0.37 (95% CI: 0.21, 0.75).

Employing countermatching to improve the efficiency of this analysis with a small number of controls per case, we conducted a 1:3 matched study with countermatching on the radon exposure distribution of cases (15). The estimated ERR/100 WLM was 0.40 (95% CI: 0.22, 0.81)—very similar in magnitude to the estimate we obtained via a simple random sample of controls in our 1:2 matched analysis but with a confidence interval width closer to that obtained via the 1:40 matched analysis. Next, we fitted a model that adjusted for cumulative smoking. The model was of the form $\varphi(R(t), S(t); \beta_R, \beta_S) = (1 + \beta_R R(t))(1 + \beta_S S(t))$, where $R(t)$ and $S(t)$ are cumulative radon and smoking 2 years prior to age t, respectively. The smoking-adjusted estimate of β_R was ERR/100 WLM = 0.38 (95% CI: 0.21, 0.77).

Lastly, we conducted a Cox regression analysis of the full cohort by sampling all controls from the risk set for each case. Fitting an ERR model for cumulative exposure, adjusted for smoking, yielded an estimated ERR/100 WLM of 0.38 (95% CI: 0.22, 0.75), very similar to the countermatched data analysis. When the ties were not broken, there were 28 risk sets with 2 cases. In this example, the Breslow likelihood and multiple-case conditional logistic regression-estimated smoking-adjusted cumulative radon ERR/100 WLM were the same as the single-case risk-set estimates to 2 decimal places.

DISCUSSION

Over the last 2 decades, the methods of Cox regression and conditional logistic regression analysis of matched case-control data have become widely used. These analytical methods are now routinely introduced in intermediate courses in epidemiology and other disciplines and are readily implemented using standard statistical packages. However, most standard statistical packages restrict the data analyst to use of log-linear model forms. In this paper, we have described an approach to using the SAS statistical package to fit general relative risk models via the conditional logistic likelihood. This permits the fitting of Cox models and conditional logistic regression models to matched case-control data in settings in which the investigator wishes to use non-log-linear model forms. Approaches to using SAS to fit Poisson and unconditional logistic general relative risk models have been described previously (5, 6). However, to our knowledge, the Cox model and the conditional logistic regression model have not been addressed in the prior literature.

While investigators can often accommodate non-log-linearity by fitting more complex models, when effects diverge from the log-linear form it is often desirable and more informative to summarize the dose-response effect as linear on another scale or in some other way altogether, especially when there is a practical reason or biologic basis. However, non-log-linear model forms, such as the linear ERR model, typically have computational restrictions, since the relative risk cannot be negative. Consequently, the point estimate and/or confidence limits for a parameter may not be obtained. In some of our examples, the estimates could be obtained but the range of valid parameter values needed to be restricted via a bounds statement. In contrast, log-linear models have the desirable property that estimated hazard rates or odds are necessarily positive quantities, regardless of the values of the linear predictor in the regression model.

General relative risk models allow for alternatives to the usual assumption of an exponential dependence of relative risk on exposure variables and multiplicative interactions. Rather, exposure-response associations can be modeled under a wide variety of parametric forms. General relative risk models may be of particular interest in epidemiologic studies that focus on characterizing the form of a dose-response association or the nature of interactions between model covariates. Historically, one obstacle to fitting general relative risk models has been implementation using standard statistical packages. Analysts have tended to use specialized software or FORTRAN code that was written specifically for fitting models of this form to epidemiologic data (7, 26). In this paper, we have illustrated how general relative risk models for survival time and matched case-control data may be fitted using a standard statistical package. We have used SAS software to illustrate the methods, but we hope that they will be implemented using other software packages as well. This paper should facilitate investigation of alternatives to log-linear models in epidemiologic data.

ACKNOWLEDGMENTS

Author affiliations: Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California (Bryan Langholz); and Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, North Carolina (David B. Richardson).

Conflict of interest: none declared.

REFERENCES

1. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Ser B*. 1972;34(2):187–220.
2. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2007.
3. Robins JM, Gail MH, Lubin JH. More on ‘‘Biased selection of controls for case-control analyses of cohort studies.’’ *Biometrics*. 1986;42(2):293–299.
4. Greenland S. Modeling and variable selection in epidemiologic analysis. *Am J Public Health*. 1989;79(3):340–349.
5. Richardson DB. A simple approach for fitting linear relative rate models in SAS. *Am J Epidemiol*. 2008;168(11): 1333–1338.
6. Richardson DB, Kaufman JS. Estimation of the relative excess risk due to interaction and associated confidence bounds. *Am J Epidemiol*. 2009;169(6):756–760.
7. Thomas D. General relative risk models for survival time and matched case-control analysis. *Biometrics*. 1981;37(4): 673–686.
8. Lubin JH. Models for the analysis of radon-exposed populations. *Yale J Biol Med*. 1988;61(3):195–214.
9. Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford University Press; 2003.
10. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text*. New York, NY: Springer Publishing Company; 2005.
11. Richardson DB. An incidence density sampling program for nested case-control analyses [electronic article]. *Occup Environ Med*. 2004;61(12):e59.
12. Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect.* 1994;102(suppl 8): 47–51.
13. Langholz B, Borgan Ø. Counter-matching: a stratified nested case-control sampling method. *Biometrika*. 1995;82(1):69–79.
14. Borgan Ø, Goldstein L, Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann Stat*. 1995;23(6):1749–1778.
15. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Stat Sci.* 1996;11(1):35–53.
16. Langholz B, Goldstein L. Conditional logistic analysis of case-control studies with complex sampling. *Biostatistics*. 2001; 2(1):63–84.
17. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73(1):1–11.
18. Binder D. Fitting Cox's proportional hazards models from survey data. *Biometrika*. 1992;79(1):139–147.

19. Borgan O, Langholz B, Samuelsen SO, et al. Exposure stratified case-cohort designs. *Lifetime Data Anal*. 2000;6(1): 39–58.
20. Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. *Scand J Stat*. 2007;34(1):103–119.
21. Breslow N. Covariance analysis of censored survival data. *Biometrics*. 1974;30(1):89–99.
22. Allison PD. *Survival Analysis Using SAS: A Practical Guide*. Cary, NC: SAS Institute Inc; 1995.
23. Gail MH, Lubin J, Rubenstein LV. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*. 1981;68(3):703–707.
24. Hornung RW, Meinhardt TJ. Quantitative risk assessment of lung cancer in U.S. uranium miners. *Health Phys.* 1987;52(4): 417–430.
25. SAS Institute Inc. *SAS OnlineDoc 9.2*. Cary, NC: SAS Institute Inc; 2007. (http://support.sas.com/documentation/). (Accessed November 11, 2009).
26. Preston DL, Lubin JH, Pierce DA, et al. *Epicure: User's Guide*. Seattle, WA: HiroSoft International Corporation; 1993.

APPENDIX 1

Creating a Case-Control-Level Data Structure From a Person-Level Data Structure

The case-control-level data structure (cclevel) shown in Table 1 can be generated from the person-level data structure (plevel) via a simple SAS data step, as shown below.

```
data cclevel (keep= setno z1-z3);
  set plevel; by setno;
  array z{3};
  retain i z1-z3;
  if first.setno then i = 0;
  i = i + 1;
  z{i} = z;
  if last.setno then output;
  run;
```

More generally, a SAS macro "make_case_control.sas" for creating a case-control data structure from a person-level data set is available at http://hydra.usc.edu/timefactors/examples/exampl.html (topic 10). The macro facilitates the handling of data sets with large numbers of explanatory variables, as well as variable numbers of cases and controls.

Also available at http://hydra.usc.edu/timefactors is a SAS macro for creating risk-set data from cohort data, as well as example data sets and SAS code for performing the analyses described in this paper.

APPENDIX 2

Fitting Non-Log-Linear Models to 1:Variable Matched Case-Control Data

Consider a case-control-level data structure in which the largest case-control sets have 11 members and the variable *ntot* gives the number of subjects in the current set. Subject 1 is the case. A conditional logistic regression analysis of these data employing the linear model form $\varphi(z;\beta) = (1 + \beta z)$ could be fitted via SAS as follows:

```
proc nlp data=;
  parms beta=0;
  profile beta / alpha=.05;
  array z z1-z11;
    sum=0; ntot=ntot;
    do i = ntot to 1 by -1;
    phi= 1+(z{i}*beta);
    sum=sum + phi;
    end;
  L=log(phi/sum);
  max L; run;
```

The profile statement specifies that 95% profile likelihood confidence intervals are to be output. The ntot = ntot line is needed to initialize the variable. A mixture model of the form $\varphi(z;\beta) = (e^{z\beta})^{\alpha}(1 + z\beta)^{1-\alpha}$ may be fitted to 1:variable matched case-control data via SAS, as follows:

```
proc nlp data= ;
  parms beta=0, alpha=0;
  profile alpha / alpha=.05;
  array z z1-z11;
  sum=0; ntot=ntot;
  do i = ntot to 1 by -1;
    phi=((exp(z{i}*beta)**alpha)*((1+z{i}*beta)**(1-
        alpha)));
    sum=sum + phi;
    end;
  L=log(phi / sum);
  max L; run;
```

Note that the parms statement now includes 2 parameters (β and α), both of which are initialized to 0.

APPENDIX 3

Fitting Non-Log-Linear Models to Risk-Set Data With Countermatching

The largest case-control sets have 11 members, and the variable *ntot* gives the number of subjects in the current set. The countermatching weights are given by the variables $w1$–$w11$. A model of the form $\varphi(z;\beta) = e^{(z\beta)}$can be fitted to countermatched data, using the following SAS code.

```
proc nlp data=;
  parms beta=0;
  array z z1-z11;
  array w w1-w11
  sum=0; ntot=ntot;
  do i = ntot to 1 by -1;
```

```
    phi= exp(z{i}*beta)*w{i};
    sum=sum + phi;
    end;
  L=log(phi/sum);
  max L; run;
```

APPENDIX 4

Fitting Models to Risk-Set Data With Multiple Cases in the Case-Control Set Using the Recursive Algorithm

As an example, a model of the form $\varphi(z;\beta) = e^{(z\beta)}$ can be fitted to *d*:*m* data with a maximum of 5 cases and 11 subjects in the case-control set using the following code:

```
proc nlp data=;
  parms beta=;
  array z1-z11;
  array b0a{5} b0a1-b0a5;
  array b0b{5} b0b1-b0b5;

* calculate denominator using the recursive formula;
* at the end b0b{ncases} is the sum of products;

* initialize k-1 level and compute case set or;
  casesetor = 1; ncases=ncases;
  do i = 1 to ncases;
    phi = exp(z{i}*beta);
    casesetor = casesetor*phi;
    if i eq 1 then b0a{i} = phi; else b0a{i} = 0;
    end;

* do the recursion;
  ntot = ntot;
  do i = 2 to ntot;
    phi = exp(z{i}*beta);
    b0b{1} = b0a{1} + phi;
      do j = 2 to min(i,ncases);
        b0b{j} = b0a{j} + b0a{j-1}*phi;
        end;

* re-initialize the k-1 step array;
    do j = 1 to min(i,ncases);
      b0a{j} = b0b{j};
      end;
    end;

* log partial likelihood contribution from the risk set;
    L = log(casesetor / b0b{ncases});
    max L; run;
```